Обучение машинного перевода без параллельных текстов *

Гончаров*[1] *М. Ю., Бахтеев*[1] *О. Ю., Стрижов*[2] *В. В.
goncharov.myu@phystech.edu
[1]Московский физико-технический институт
[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

В данной работе исследуется метод обучения без учителя для решения задачи машинного перевода. Метод основан на нейросетевой модели seq2seq, в которой кодировщик отображает предложение в вектор латентного пространства, а декодировщик, используя этот вектор, восстанавливает исходное предложение на любом из двух языков. Процесс оптимизации параметров нейросети построен таким образом, чтобы избежать необходимости использовать параллельные данные. Качество работы метода проверяется на паре языков русский-украинский.

Ключевые слова: *машинный перевод, seq2seq, обучение без учителя.*

Введение

Задача машинного перевода заключается в том, чтобы автоматически переводить текст с одного языка на другой. Существует ряд подходов к решению этой задачи. Идея статистического подхода состоит в том, чтобы рассматривать задачу перевода как задачу машинного обучения. Перевод предложения рассматривается как случайная величина с неизвестным распределением, которое моделируется распределением из некоторого параметрического семейства, параметры которого можно найти методом наибольшего правдоподобия при наличии выборки из параллельных предложений. Нейросетевой машинный перевод использует для аппроксимации неизвестного распределения нейросетевые модели, которые чаще всего состоят из кодировщика и декодировщика, которые в большинстве случаев представляют из себя реккурентные нейронные сети [1, 2]. Кодировщик получает на вход предложение на одном из языков и отображает его в вектор латентного пространства. Декодировщик, получая на вход этот вектор, пошагово восстанавливает исходное предложение на другом языке. Такая модель получила название seq2seq. В последние годы, после появления механизма attention, с помощью модели seq2seq были достигнуты существенные результаты для пар языков, для которых есть достаточное количество параллельных данных (порядка миллиона предложений) [3, 4].

К сожалению, для некоторых пар языков, например когда один из них относительно мало используется, собрать такое количество качественных размеченных данных может быть сложно или невозможно. Поэтому были предложены методы, позволяющие полностью или частично отказаться от использования параллельных данных. Одним из них является метод, предложенный в [5]. Его идея заключается в следующем. Предположим, что мы имеем выборку из качественных предложений на одном из языков. Тогда при оптимизации параметров нейросети будем использовать предложения из выборки, в качестве целевой переменной, а на вход подавать их некачественные переводы, полученные с помощью простого переводчика, например обученного без учителя пословного переводчика

Научный руководитель: Стрижов В. В. Авторы: А.В. Грабовой, О.Ю. Бахтеев, В.В. Стрижов, Eric Gaussier, координатор Малиновский Г. С. Консультант: Бахтеев О. Ю.

[6]. Таким образом, нейросеть будет решать задачу обратного перевода с исправлением ошибок.

В данной работе изложены детали этого метода и проводится вычислительный эксперимент, в ходе которого проверяется применимость метода к паре языков русский-украинский. Качество перевода оценивается с помощью метрики BLEU [7].

Постановка задачи

Даны обучающая выборка $\mathfrak{D}^{\text{src}} = \{\mathbf{s}_1^{\text{src}}, \dots, \mathbf{s}_{m_{\text{src}}}^{\text{src}}\}$, $\mathfrak{D}^{\text{tgt}} = \{\mathbf{s}_1^{\text{tgt}}, \dots, \mathbf{s}_{m_{\text{tgt}}}^{\text{tgt}}\}$, которая состоит из двух корпусов произвольных предложений для каждого языка; и валидационная выборка $\mathfrak{D}^{valid} = \{(\tilde{\mathbf{s}}_1^{\text{src}}, \tilde{\mathbf{s}}_1^{\text{tgt}}), \dots, (\tilde{\mathbf{s}}_{m_{\text{valid}}}^{\text{src}}, \tilde{\mathbf{s}}_{m_{\text{valid}}}^{\text{tgt}})\}$, которая состоит из пар параллельных предложений.

Задача состоит в том, чтобы, используя обучающую выборку, построить модель M перевода предложений с языка src на язык tgt и выбрать ее параметры таким образом, чтобы максимизировать среднее значение метрики BLEU на валидационной выборке

$$\frac{1}{m_{\text{valid}}} \sum_{i=1}^{m_{\text{valid}}} \text{BLEU}(M(\tilde{\mathbf{s}}_i^{\text{src}}), \tilde{\mathbf{s}}_i^{\text{tgt}}) \to \max_M$$

Гипотеза

Пусть $\mathcal{S}^l$ – множество всех предложений на языке $l \in \{\text{src}, \text{tgt}\}$. Выдвигается гипотеза о том, что существует единое латентное пространство $\mathcal{L}$ и отображения $\boldsymbol{f}^l : \mathcal{S}^l \to \mathcal{L}$ и $\boldsymbol{g}^l : \mathcal{L} \to \mathcal{S}^l$ такие, что

— для $\forall l_1, l_2 \in \{\text{src}, \text{tgt}\}$ и $\forall s^{l_1} \in \mathcal{S}^{l_1}$ $\boldsymbol{g}^{l_2}(\boldsymbol{f}^{l_1}(s^{l_1}))$ совпадает с s^{l_1}, если l_1 совпадает с l_2, и является корректным переводом s^{l_1}, если l_1 и l_2 различаются

— распределения образов $\boldsymbol{f}^{\text{src}}(\mathcal{S}^{\text{src}})$ и $\boldsymbol{f}^{\text{tgt}}(\mathcal{S}^{\text{tgt}})$ совпадают

Описание метода

Предлагаемый метод заключается в том, чтобы для каждого $l \in \{\text{src}, \text{tgt}\}$ моделировать отображения $\boldsymbol{f}^l$ и $\boldsymbol{g}^l$ кодировщиком $\mathbf{f}^l$ и декодировщиком $\mathbf{g}^l$ соответственно. Таким образом, моделью перевода является композиция $\mathbf{g}^{\text{tgt}} \circ \mathbf{f}^{\text{src}}$.

Оптимизация проводится следующим образом. Функция ошибки содержит три слагаемых, которые соответствуют сделанным предположениям об отображениях $\boldsymbol{f}^l$ и $\boldsymbol{g}^l$.

Ошибка восстановления Для каждого $l \in \{\text{src}, \text{tgt}\}$ рассматривается входное предложение $\mathbf{s}^l$ на языке l. Ошибки считаются между $\mathbf{s}^l$ и его образом при отображении $\mathbf{g}^l \circ \mathbf{f}^l$.

Ошибка перевода Рассматривается предложение $\mathbf{s}^{\text{src}}$ без ограничения общности на языке src. В качестве входного предложения используется $\mathbf{s}^{\text{tgt}}$ – перевод $\mathbf{s}^{\text{src}}$, полученный с помощью некоторой слабой модели перевода M^0. Ошибка перевода с языка tgt на язык src считается между $\mathbf{s}^{\text{src}}$ и образом $\mathbf{s}^{\text{tgt}}$ при отображении $\mathbf{g}^{\text{src}} \circ \mathbf{f}^{\text{tgt}}$. Аналогичным образом считается ошибка перевода с языка src на язык tgt. В качестве M^0 используется пословный переводчик на основе предобученных векторных представлений слов $\mathcal{E}^{\text{src}} = \{\mathbf{x}_i\}_{i=1}^{n_{\text{src}}}$ и $\mathcal{E}^{tgt} = \{\mathbf{y}_i\}_{i=1}^{n_{\text{tgt}}}$. Переводом слова $\mathbf{x} \in \mathcal{E}^{\text{src}}$ является $\mathbf{y} = \arg\min_{\mathbf{y} \in \mathcal{E}^{\text{tgt}}} \rho(\mathbf{x}, \mathbf{y})$, где ρ – косинусное расстояние (перевод слов языка tgt осуществляется аналогично).

Штраф за различие распределений Вводится дискриминатор $\mathbf{d}$, который решает задачу классификации векторов $\mathbf{h}^l$ латентного пространства $\mathcal{L}$ на классы $\{0, 1\}$: $\mathbf{h}^l \in 0 \Leftrightarrow l = \text{src}$. Векторы $\mathbf{h}^l$ получаются кодировщиком $\mathbf{f}^l$, в качестве промежуточного результата при отображениях входных предложений. В функцию ошибки добавляется

штраф за то, что дискриминатор точно определяет язык входного предложения. Таким образом параметры модели оптимизируются таким образом, чтобы усложнить задачу дискриминатору. В свою очередь параметры дискриминатора оптимизируются параллельно параметрам модели. Соревновательный процесс оптимизации мотивирован желанием добиться сходства распределений латентных векторов для разных языков.

Детали метода

Введем словарь V^{src}, содержащий проиндексированные слова, встречающиеся в предложениях из $\mathfrak{D}^{\text{src}}$ и переводах предложений из $\mathfrak{D}^{\text{tgt}}$, получаемых с помощью M^0. Аналогично введем V^{tgt}.

Предложения на разных языках, использующиеся при оптимизации кодируются с помощью соответствующих словарей. Входные предложения зашумляются преобразованием σ: сначала с некоторой вероятностью q из них удаляется каждое слово, а затем производится случайная перестановка оставшихся слов с условием, что слово не может оказаться дальше чем на k позиций от своей начальной позиции (q, k – гиперпараметры).

Кодировщик $\mathbf{f}^l$ включает в себя слой векторного представления слов $\mathbf{e}^l$, параметры которого инициализируются векторами $\mathcal{E}^l$, и реккурентную нейронную сеть $\mathbf{r}^{\text{enc}}$. Декодировщик $\mathbf{g}^l$ включает в себя $\mathbf{e}^l$, $\mathbf{r}^{\text{dec}}$ и классификатор $\mathbf{c}^l$, который решает задачу многоклассовой классификации выходов $\mathbf{r}^{\text{dec}}$ на классы, соотвествующие словам в словаре V^l (отображает выходы $\mathbf{r}^{\text{dec}}$ в векторы вероятностей размерности $|V^l|$). Параметры $\mathbf{r}^{\text{enc}}$ и $\mathbf{r}^{\text{dec}}$ общие для $\mathbf{f}^{\text{src}}$, $\mathbf{f}^{\text{tgt}}$ и $\mathbf{g}^{\text{src}}$, $\mathbf{g}^{\text{tgt}}$ соответственно.

В качестве меры ошибок классификации используется кросс-энтропия CE. Итоговый вид функции ошибки

$$L_{\text{tran}} = w_1 \cdot \sum_{l \in \{src,tgt\}} CE(\mathbf{g}^l(\mathbf{f}^l(\sigma(\mathbf{s}^l))), \mathbf{s}^l) +$$

$$+ w_2 \cdot \sum_{l_1 \neq l_2 \in \{src,tgt\}} CE(\mathbf{g}^{l_2}(\mathbf{f}^{l_1}(\sigma(\mathbf{s}^{l_1}))), \mathbf{s}^{l2}) + w_3 \cdot \sum_{l \in \{src,tgt\}} CE(\mathbf{d}(\mathbf{f}^l(\sigma(\mathbf{s}^l))), \mathbb{I}[l = \text{src}])$$

$$L_{\text{disc}} = \sum_{l \in \{src,tgt\}} CE(\mathbf{d}(\mathbf{f}^l(\sigma(\mathbf{s}^l))), \mathbb{I}[l = \text{tgt}]).$$

Литература

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[2] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[4] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[5] G. Lample, L. Denoyer, and M. Ranzato, "Unsupervised machine translation using monolingual corpora only," *arXiv preprint arXiv:1711.00043*, 2017.

[6] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *arXiv preprint arXiv:1710.04087*, 2017.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318, Association for Computational Linguistics, 2002.